



Fundamentos de Big Data

Lección 1: Fundamentos básicos teóricos de Big Data

ÍNDICE

Lección 1. – Fundamentos básicos teóricos de Big Data	2
Presentación y objetivos	2
1. Contenidos de la Asignatura	4
2. Fundamentos básicos: introducción	6
3. Cambio de Paradigma en las Organizaciones	8
4. Ejemplos Reales de Casos de Uso Big Data	9
5. El Gobierno del Dato / Gobernanza de datos	11
6. Analítica Avanzada de Datos.....	12
7. Tecnologías catalizadoras del Big Data	13
8. Herramientas de Visualización.....	13
9. ¿ Qué es Kaggle ? ¿ Qué es el Titanic Dataset ?	14
10. Nos Registramos en Kaggle	15
11. Puntos clave	21

Lección 1. – Fundamentos básicos teóricos de Big Data

PRESENTACIÓN Y OBJETIVOS

Esta asignatura contiene una gran cantidad de elementos teóricos, que serán sintetizados para poder abordar la parte más importante para cualquier programador/a que es el propio código.

Tratará de hacer una visión general de las herramientas más interesantes que existen a fecha mayo 2021.

Y, tal y como se encuentra el Big Data a día de hoy es posible que salgan cosas nuevas (y quizá mejores) dentro de poco tiempo.



Objetivos

- Conocer de manera teórica los conceptos básicos fundamentales del Big Data
- Conocer algunas de las mejores herramientas para realizar Gráficas
- Conocer Kaggle como herramienta de aprendizaje en Data Science



Fuentes de obtención de los Logos:

<https://www.python.org/community/logos/>

<https://www.kaggle.com/arunsankar/kaggle-logo>

<https://www.vectorlogo.zone/>

<https://github.com/scikit-learn/scikit-learn/tree/main/doc/logos>

<https://pandas.pydata.org/about/citing.html>

<https://numpy.org/>

<https://github.com/numpy/numpy/blob/main/branding/logo/primary/numpylogo.png>

https://commons.wikimedia.org/wiki/File:Jupyter_logo.svg

<https://bokeh.org/branding/>

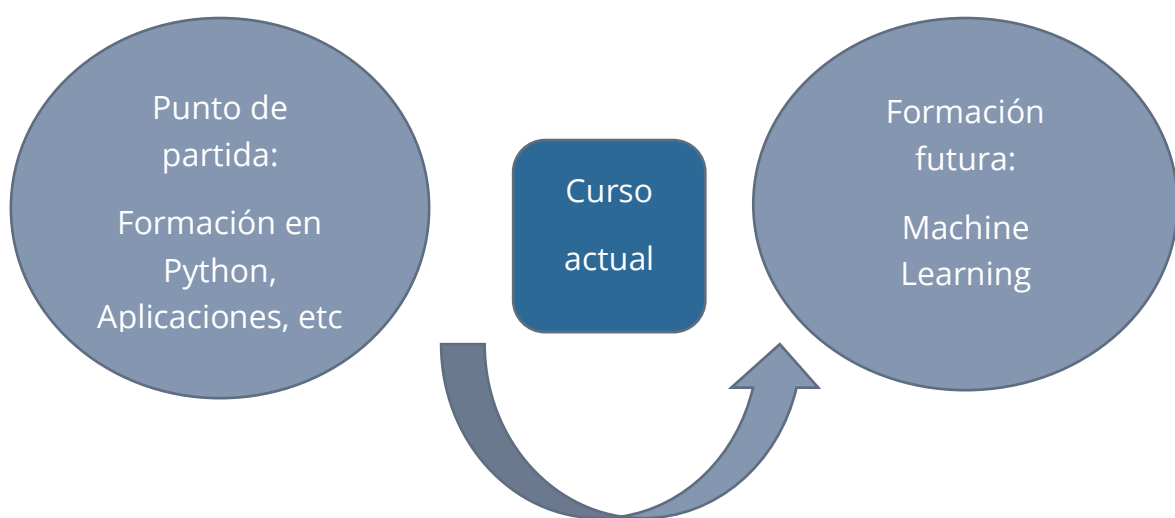
<https://plotly.com/dash/>

<https://seaborn.pydata.org/citing.html>

https://github.com/mwaskom/seaborn/blob/master/doc/_static/logo-wide-lightbg.png

1. CONTENIDOS DE LA ASIGNATURA

Explicación del contenido global del curso y de los motivos de estar así enfocado



Se hace difícil explicar Big Data sin explicar algo de Machine Learning, por lo que no entraremos al detalle del todo, pero explicaremos algunas cosas.

De igual manera, tomaremos como aprendido lo visto en "Creación de Aplicaciones Python".

Hay otra materia de Big Data, por lo que se tratará de abordar en esta asignatura algunos de los conocimientos, pero para entrar en un poco más en profundidad será necesario cursar ambas materias.

Para ello, y en esta primera asignatura hemos diseñado el siguiente itinerario formativo:

Fundamentos de Big Data

Tema 1: Fundamentos básicos teóricos en Big Data

Tema 2: Librerías para Gráficas: Pandas, Matplotlib, Seaborn..

Tema 3: Principales Tipos de Gráficas: Histogramas, Pie Charts..

Tema 4: Introducción a “Data Mining” con Titanic Dataset

Tema 5: Kaggle y los retos de Data Science

Existen muchas más alternativas, pero será algo que podría dar para más horas.

2. FUNDAMENTOS BÁSICOS: INTRODUCCIÓN

Primeramente comentar que lo mencionado en este punto no es algo fijo, sino que, cada empresa organiza los puestos y responsabilidades de una forma, pero, en esencia, y por sintetizar se comentará lo siguiente.

Existen varios “roles” en Inteligencia Artificial y Big Data.

Por sintetizar (aunque puede haber variaciones):

- **Perfil “Data Engineer”:**

Big Data Engineer

Es la persona que se encarga de la Arquitectura.

Tecnologías: Bases de Datos SQL y NoSQL

Apache Spark (PySpark y/o Scala)

Hadoop

Conocimientos ETL

- **Perfil “Data Scientist”:**

Conocimiento de los Algoritmos de Machine Learning.

Conocimiento de Herramientas de Visualización de datos.

Bases de datos SQL y NoSQL

A veces existe un perfil de Machine Learning / Deep Learning específico y los “Data Scientist” son perfiles con menos conocimiento en esto y más en Visualización.

- También existe un **perfil “Data Analyst”** que hace básicamente lo que un Data Scientist, pero del cual se espera (teóricamente) menos conocimiento, especialmente en la parte de Algoritmos y redes neuronales, podría ser alguien que haga muchas gráficas.

En la práctica lo ideal es tener perfiles híbridos que aunque se especialicen en una parte comprendan todas las etapas. Además, este listado no es del todo completo puesto que hay más tipos de profesiones ligados a este campo.

Tras la obtención de los datos, es necesario almacenarlos.

Existen varios tipos de proyectos (entre los cuales mencionamos):

- Regresión,
- Clasificación,
- Clustering,

Lo 1ª será indicar QUÉ necesitamos hacer.

Y una vez lo tengamos claro necesitamos preparar la información para ser tratada de esa forma.

A continuación deberíamos hacer lo que se conoce como “limpieza de datos” o data cleaning. Existen unas cuantas etapas en el preprocesamiento, las cuales se verán posteriormente.

Las etapas de predicción son las finales, y obviamente, evaluación del modelo.

Es posible que necesitemos el uso del Cloud para analizar.

En cuanto a las gráficas son cosas que se hacen en varios puntos, como “Análisis de datos exploratorio” o al final del proyecto como conclusiones o visualización de los clústeres, por ejemplo.

3. CAMBIO DE PARADIGMA EN LAS ORGANIZACIONES

Fruto de la existencia de Tecnologías relacionadas con la Inteligencia Artificial (IA) se está llevando a cabo un cambio total en la forma que las empresas toman sus decisiones.

Se plantean:

- ✓ ¿Y si podemos usar IA, para mejorar los procesos de fabricación ?
(con técnicas de visión artificial, Deep Learning, etc.
- ✓ ¿Y si podemos sacar provecho a todos los datos recopilados desde hace 20 años para tomar decisiones ?
Ahora es posible procesar grandes volúmenes de datos relativamente rápido.
- ✓ ¿Y si podemos hacer análisis de la competencia en base a los datos ?
Obteniendo datos de webs, y procesando.

Existen cosas que llevan haciéndose muchísimos años, pero si bien es cierto que ahora tenemos la tecnología más al alcance.

4. EJEMPLOS REALES DE CASOS DE USO BIG DATA

Ejemplos reales de casos de uso de Inteligencia Artificial y Big Data:

- Predicción de precios en la bolsa (Trading automático)
https://en.wikipedia.org/wiki/Algorithmic_trading
- Big Data para Supply Chain Management (SCM)
<https://www.mckinsey.com/business-functions/operations/our-insights/big-data-and-the-supply-chain-the-big-supply-chain-analytics-landscape-part-1#>

Big Data en Fabricación

<https://www.ibm.com/downloads/cas/ONBGKB82>

<https://www.mckinsey.com/business-functions/operations/our-insights/how-big-data-can-improve-manufacturing>

- Sistemas de detección de fraude (bancario por ejemplo)
<https://www.kaggle.com/mlg-ulb/creditcardfraud>
- Sistemas de recomendación
Ejemplo recomendaciones en Netflix, o YouTube, etc.
Para el caso de Netflix:
<https://help.netflix.com/en/node/100639>
- Vehículos de conducción autónoma
El Autopilot de Tesla
<https://www.tesla.com/autopilot>
- Fútbol y baloncesto, análisis tácticos, posiciones de los jugadores en el campo (para estrategias..).
Liga Inglesa de Fútbol:
<https://www.dailymail.co.uk/sport/football/article-9392769/Football-big-clubs-using-data-analysis-edge-rivals-technology-war.html>

- Atlético de Madrid (fútbol):
<https://www.mundodeportivo.com/futbol/atletico-madrid/20180209/44637421202/analisis-big-data-secreto-exito-del-atletico-madrid.html>
- NBA: Proyecto de la Harvard Business School
<https://digital.hbs.edu/platform-digit/submission/how-data-analytics-is-revolutionizing-the-nba/>
- Negociación de contratos en el fútbol sin representante
<https://www.infobae.com/america/deportes/2021/04/08/sin-representante-kevin-de-bruyne-utilizo-un-novedoso-metodo-para-firmar-una-renovacion-millonaria-con-el-manchester-city/#:~:text=Deportes-,Sin%20representante%2C%20Kevin%20De%20Bruyne%20utiliz%C3%B3%20un%20novedoso%20m%C3%A9todo%20para,millonaria%20con%20el%20Manchester%20City>
- Big Data en el Pádel
<https://www.bullpadel.com/es/blog/nito-brea-el-analisis-estadistico-en-el-padel--n265>
- Carolina Marín, Campeona del mundo varias veces (Bádminton)
<https://elpais.com/eps/2021-04-04/carolina-marin-la-reina-del-big-data.html>

Y un sinfín más de aplicaciones.

Y todavía no ha llegado la revolución del Internet of Things (IoT), que permitirá tener toda la ciudad conectada a un vehículo, o la nevera de casa al supermercado, quién sabrá antes que tú cuando necesitas comprar yogures.

5. EL GOBIERNO DEL DATO / GOBERNANZA DE DATOS

La gobernanza de datos (o en inglés “Data governance”) es la gestión de datos de una empresa.

Trata de ver a los datos desde todos los puntos de vista posibles.

Tiene, entre otras, las siguientes características:

- Debe asegurar el cumplir con normativas
- Debe ayudar a reducir costes
- Debe aprovechar los datos como uno de elementos más importantes de una empresa
- Se debe disponer de la información en el momento correcto, no tarde, y con el formato adecuado.
- Debe haber seguridad de los datos, es decir, que solo accedan a ellos las personas autorizadas.
- Los datos deben ser capaces de agregar valor
- Los datos son necesarios en tareas de Business Intelligence (BI)
- Existen auditorías o revisiones de calidad (se mide como cualquier otra cosa en una empresa)
- Hay una Gestión de Datos (Data Management)
- En la empresa hay diferentes roles que colaboran en ello.

“Data Steward” es uno de ellos (aunque no siempre se contrata a una persona para hacer exclusivamente esto, y puede haber varios)

6. ANALÍTICA AVANZADA DE DATOS

Con la mejora de la tecnología es posible tener uso de los datos históricos y usarlos en el momento adecuado para Business Intelligence (BI).

Podemos, pues, dar respuesta a:

- Qué sucedió en la empresa y por qué (Análítica Descriptiva)
En esta fase se usan herramientas de Visualización de Datos.
(Presente en base al pasado)
- Qué podría suceder “mañana” en base a “ayer” (Análítica Predictiva)
Para ello se usan técnicas de matemáticas y estadística para tratar de pronosticar.
(Predicción del futuro en base a históricos pasados y presente)

Proyecto de Análítica Avanzada de Datos:

- **1ª Etapa:** Evaluar las necesidades que se tienen.
- **2ª Etapa:** Recopilar la información (crear un “Data Lake”)
- **3ª Etapa:** Verificación de la “Calidad” de los datos, protección, privacidad, seguridad
- **4ª Etapa:** Se mide todo, como en cualquier proyecto. A nivel gerencial se usan los Key Performance Indicators (KPIs), o indicadores de calidad. Estos indicadores son fijados antes de comenzar, obviamente

7. TECNOLOGÍAS CATALIZADORAS DEL BIG DATA

Algunas de estas Tecnologías Big Data (y algunas cosas más) serán vistas en la 2ª Asignatura de Big Data:

- Hadoop
- Apache Spark
- MongoDB (NoSQL-Bases de datos No relacionales)
- Cassandra DB (NoSQL-Bases de datos No relacionales)
- Apache Airflow
- Docker
- Kubernetes
- TensorFlow 2.0 (Inteligencia Artificial con Keras y TF).
- Tableau (visualización de gráficos)
- Dash (visualización de gráficos)

8. HERRAMIENTAS DE VISUALIZACIÓN

Tal y como hemos señalado:

- PowerBI,
- Tableau
- Dash,

Etc

Sobre estas herramientas hablaremos en alguna lección de esta asignatura.

9. ¿ QUÉ ES KAGGLE ? ¿ QUÉ ES EL TITANIC DATASET ?

El Titanic Dataset es otro de los Datasets iniciales con los que se trabaja en Data Science (al igual que el Iris Dataset).

En este caso, y a diferencia del Iris Dataset, este Dataset aunque no muy complejo, tiene una serie de peculiaridades que lo hacen más difícil que el anterior.

Es lo que se conoce como un set de datos “sucio”, puesto que hay que modificar algunas columnas usando Natural Language Processing (NLP) o lo que es lo mismo, Procesamiento del Lenguaje Natural.

Sin lugar a duda, una buena forma de aprender Data Science y/o practicar conceptos es Kaggle, un lugar donde se plantean retos de data Science, algunos de ellos con succulentos premios.

Kaggle, desde 2017, pertenece a Google, además, tal y como se comenta en la siguiente noticia:

https://techcrunch.com/2017/03/08/google-confirms-its-acquisition-of-data-science-community-kaggle/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLnNvbS8&guce_referrer_sig=AQAAAlf2yVP95bu7rmzTzTJclp6cSo_CYCB46SBNr6wGLJxvLE_upEWxeU1rTLTPO3Mghn24JZLFjElvylvY36wg7neD92aughd6alUymmK4OOHP2aX2r4SH88qyqE-k3lIM1j2u6eV1Od2U6A4na924DN6NNU6w_AkWPCJ_Fo8DB5_j

El Titanic Dataset, de hecho le obtenemos de esta web en el siguiente enlace.

<https://www.kaggle.com/c/titanic>

El cual tendremos la oportunidad de trabajar con el mismo en las últimas lecciones de esta asignatura.

10. NOS REGISTRAMOS EN KAGGLE

Para comenzar, lo primero que haremos será registrarnos en la web de Kaggle.

Para ello, y en ese mismo Link, nos vamos a “Register”

En mi caso, me he creado una cuenta específica para explicarlo.

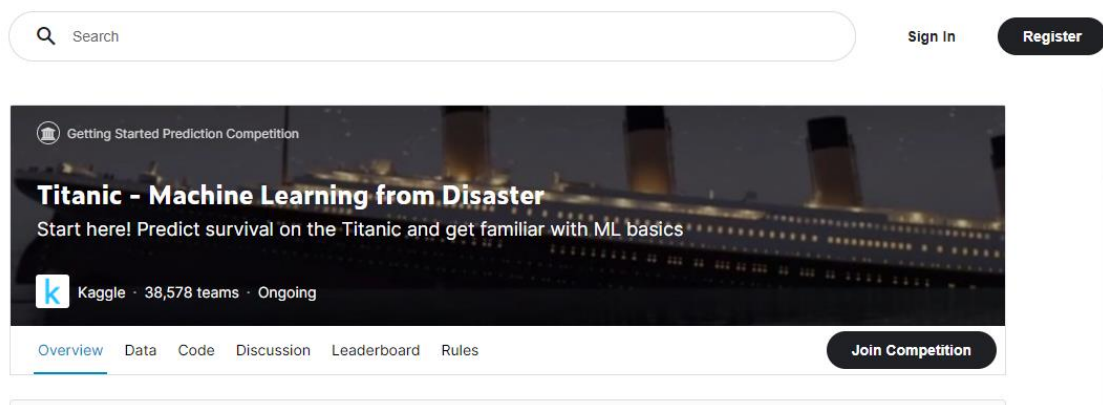




Figura 10.1: Register en Kaggle (parte 1)

Un ejemplo sería:

[Sign In](#) [Register](#)

 Register with Google

 Register with your email

Have an account? [Sign in.](#)


When you link your Facebook, Google, or Yahoo account, Kaggle collects certain information stored in that account that you have configured to make available. By linking your accounts, you authorize Kaggle to access and use your account on the third party service in connection with your use of kaggle.com.

Figura 10.2: Register en Kaggle (parte 2)

« back


Register


Email address
jmpena@grupomainjobs.com

Password (min 7 chars)
..... 

Full name (displayed)
Jose_EIP

Your profile URL
kaggle.com/JoseEIP [edit](#)

 No soy un robot


reCAPTCHA
[Privacidad](#) - [Términos](#)

☒ Email me new ML tutorials, Kaggle news and updates.
You can opt out at any time.

[Cancel](#) [Next](#)

Figura 10.3: Register en Kaggle (parte 3)

Privacy and Terms

- We collect information about the apps, browsers, and devices you use to access our Services by using different types of technology, including cookies, clear gifs, or web beacons.

Why we process it

We process this data for the purposes described in our Privacy Policy, including to:

- Deliver our services, like administering competitions you enter or hosting datasets you upload
- Improve security by protecting against fraud and abuse
- Send you messages related to Kaggle or the activities of third parties we work with
- Conduct analytics and measurement to understand how our services are used

Cancel

I agree

Figura 10.4: Register en Kaggle (parte 4)

Verify your email

We've sent you an email with a six-digit code. Please enter it here.

Six-digit code

[Resend email](#)
[Next](#)

Figura 10.5: Register en Kaggle (parte 5)

Ponemos el código que nos envíen al correo electrónico.

Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 38,578 teams · Ongoing

[Overview](#)
[Data](#)
[Code](#)
[Discussion](#)
[Leaderboard](#)
[Rules](#)

Join Competition

Overview	
Description	<div style="display: flex; align-items: flex-start;"> <div> <p>Ahoy, welcome to Kaggle! You're in the right place.</p> <p>This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works.</p> <p>The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.</p> <p>Read on or watch the video below to explore more details. Once you're ready to start competing, click on the "Join Competition button" to create an account and gain access to the competition data. Then check out Alexis Cook's Titanic Tutorial that walks you through step by step how to make your first submission!</p> </div> </div>
Evaluation	
Frequently Asked Questions	

Figura 10.6: Register en Kaggle (parte 6)

Hacemos click en el botón de la parte superior derecha:

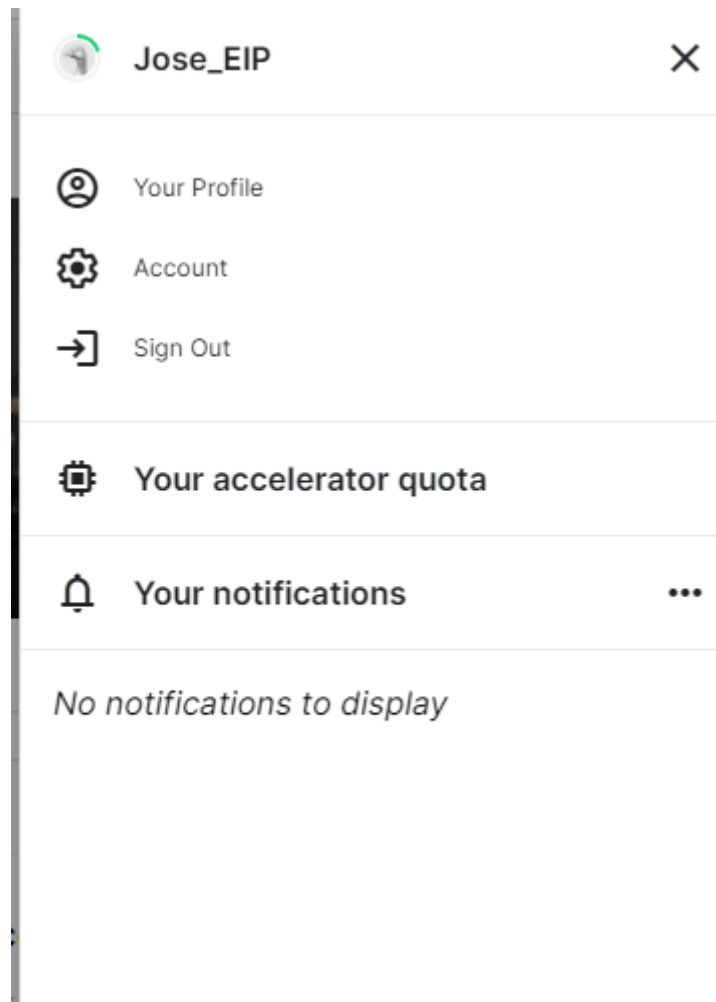


Figura 10.7: Register en Kaggle (parte 7)

Ahora podemos ver nuestro perfil ("Your Profile")

11. PUNTOS CLAVE

- | Existen muchas tecnologías que forman parte del ecosistema Big Data
- | Una parte importante en Big Data es la visualización de Gráficas para obtener información precisa de lo ocurrido en el pasado y previo a emplear técnicas avanzadas estadísticas de predicción.
- | Kaggle es una herramienta IMPRESCINDIBLE para la formación en Data Science, con retos diversos, y que nos permitirá aprender muchas cosas.

